Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models*

Ricard Marxer, Jordi Janer, and Jordi Bonada

Universitat Pompeu Fabra,
Music Technology Group,
Roc Boronat 138, Barcelona
{ricard.marxer,jordi.janer,jordi.bonada}@upf.edu

Abstract. This research focuses on the removal of the singing voice in polyphonic audio recordings under real-time constraints. It is based on time-frequency binary masks resulting from the combination of azimuth, phase difference and absolute frequency spectral bin classification and harmonic-derived masks. For the harmonic-derived masks, a pitch likelihood estimation technique based on Tikhonov regularization is proposed. A method for target instrument pitch tracking makes use of supervised timbre models. This approach runs in real-time on off-the-shelf computers with latency below 250ms. The method was compared to a state of the art Non-negative Matrix Factorization (NMF) offline technique and to the ideal binary mask separation. For the evaluation we used a dataset of multi-track versions of professional audio recordings.

Keywords: Source separation, Singing voice, Predominant pitch tracking.

1 Introduction

Audio source separation consists in retrieving one or more audio sources given a set of one or more observed signals in which the sources are mixed. In the field of music processing, it has received special attention the past few decades. A number of methods have been proposed, most of them based on time-frequency masks. We differentiate between two main strategies in the creation of the time-frequency mask depending on the constraints of the solution.

Realtime solutions are often based on binary masks, because of their simple and inexpensive computation. These solutions assume the target sources are orthogonal in the time-frequency domain. The most common binary mask used in stereo music recordings is based on panning information of the sources [15,8,13].

Non-realtime approaches do not make such an orthogonality assumption, and make use of a soft mask based on Wiener filtering [2] which requires estimating all spectrograms of the constitutive sources. For harmonic sources this estimation is often performed in two steps. First the pitch track of the target source is

* This research has been partially funded by Yamaha Corporation (Japan).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 314–321, 2012.

estimated and then the spectrum of that given pitch track is estimated. The first step often relies on melody extraction algorithms [7,6]. Some methods estimate the pitch of the components independently [10], while others perform a joint estimation of the pitches in the spectrum [10,14]. Most joint pitch estimation methods are computationally expensive since they evaluate a large number of possible pitch combinations. NMF approaches to multipitch likelihood estimation [11,5] address this pitfall by factoring the spectrogram into a multiplication of two positive matrices, a set of spectral templates and a set of time-dependent gains. In [4] and [9] the spectral templates are fixed to a set of comb filters representing the spectra generated by each individual pitch spectrum. We propose combining several sources of information for the creation of the binary mask in order to raise the quality of currently existing methods while maintaining low-latency. We propose two main sources of information for the creation of the masks. Spectral bin classification based on measures such as lateralization (panning), phase difference between channels and absolute frequency is used to create a first mask. Information gathered through a pitch-tracking system is used to create a second mask for the harmonic part of the main melody instrument.

2 Spectral Bin Classification Masks

Panning information is one of the features that have been used successfully [15,8] to separate sources in real-time. In [13] the pan and the IPD (inter-channel phase difference) features are used to classify spectral bins. An interesting feature for source separation is the actual frequency of each spectrum bin, which can be a good complement when the panning information is insufficient. Using pan and frequency descriptors we define a filter in the frequency domain using a binary mask to mute a given source:

$$m_i^{pf}[f] = \begin{cases} 0 & \text{if } p_{low} < p_i[f] < p_{high} \text{ and } f_{low} < f < f_{high}, \\ 1 & \text{otherwise.} \end{cases}$$

where $p_i[f]$ is the pan value of the spectral bin f at frame i. The parameters p_{low} and p_{high} are the pan boundaries and f_{low} and f_{high} are the frequency boundaries fixed at -0.25, 0.25 and 60Hz and 6000Hz respectively, to keep the method unsupervised.

The results show that this method produces acceptable results in some situations. The most obvious limitation being that it is not capable of isolating sources that share the same pan/frequency region. This technique is also ineffective in the presence of strong reverberation or in mono recordings which have no pan information.

3 Harmonic Mask

Harmonic mask creation is based on two assumptions: that the vocal component is fully localized in the spectral bins around the position of the singing voice

partials and that the singing voice is the only source present in these bins. Under such assumptions an optimal mask to remove the singing voice consists of zeros around the partials positions and ones elsewhere.

These assumptions are often violated. The singing voice is composed of other components than the harmonic components such as consonants, fricatives or breath. Additionally other sources may contribute significantly to the bins where the singing voice is located. This becomes clear in the results where signal decomposition methods such as Instantaneous Mixture Model (IMM) [4] that do not rely on such assumptions perform better than our binary mask proposal. However these assumptions allow us to greatly simplify the problem.

Under these assumptions we define the harmonic mask m^h to mute a given source as:

$$m_i^h[f] = \begin{cases} 0 & \text{for } (f0_i \cdot h) - L/2 < f < (f0_i \cdot h) + L/2, \forall h, \\ 1 & \text{otherwise.} \end{cases}$$

where $f0_i$ is the pitch of the i^{th} frame, and L is the width in bins to be removed around the partial position.We may also combine the harmonic and spectral bin classification masks using a logical operation by defining a new mask m_i^{pfh} as:

$$m_i^{pfh}[f] = m_i^{pf}[f] \vee m_i^h[f] \tag{1}$$

Finally, we are also able to produce a *soloing* mask $\bar{m}_i[f]$ by inverting any of the previously presented muting masks $\bar{m}_i[f] = \neg m_i[f]$.

In order to estimate the pitch contour $f0_i$ of the chosen instrument, we follow a three-step procedure: pitch likelihood estimation, timbre classification and pitch tracking.

3.1 Pitch Likelihood Estimation

The pitch likelihood estimation method proposed is a linear signal decomposition model. Similar to NMF, this method allows us to perform a joint pitch likelihood estimation. The main strengths of the presented method are low latency, implementation simplicity and robustness in multiple pitch scenarios with overlapping partials. This technique performed better than a simple harmonic summation method in our preliminary tests.

The main assumption is that the spectrum $X_i \in \mathbb{R}^{N_S \times 1}$ at a given frame i, is a linear combination of N_C elementary spectra, also named basis components. This can be expressed as $X_i = BG_i$, N_S being the size of the spectrum. $B \in \mathbb{R}^{N_S \times N_C}$ is the basis matrix, whose columns are the basis components. $G_i \in \mathbb{R}^{N_C \times 1}$ is a vector of component gains for frame i.

We set the spectra components as filter combs in the following way:

$$\varphi[m,n] = 2\pi f_l H N_P \frac{2^{\frac{iH-F/2+n}{HN_P}} - 1}{S_r \ln(2)}$$

$$B_m[k] = \sum_{n=0}^{F} w_a[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[m,n]\right) \right) e^{-j2\pi nk/N} \tag{2}$$

with $H = (1 - \alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the sample rate, $w_a[n]$ is the analysis window, N_h is the number of harmonics of our components, B_m is the spectrum of size N of the component of m^{th} pitch. Flat harmonic combs have been used in order to estimate the pitch likelihoods of different types of sources.

The condition number of the basis matrix B defined in Equation 2 is very high ($\kappa(B) \approx 3.3 \cdot 10^{16}$), possibly due to the harmonic structure and correlation between the components in our basis matrix. For this ill-posed problem we propose using the well-known Tikhonov regularization method to find an estimate of the components gains vector $\hat{G}_i$ given the spectrum X_i. This consists in the minimization of the following objective function:

$$\Phi(G_i) = |BG_i - X_i|^2 + \lambda |G_i|^2 \tag{3}$$

where λ is a positive scalar parameter that controls the effect of the regularization on the solution. Under the assumption of gaussian errors, the problem has the closed-form solution $\hat{G}_i = RX_i$ where R is defined as:

$$R = B^t[BB^t + \lambda I_{N_S}]^+ \tag{4}$$

and $[Z]^+$ denotes the MoorePenrose pseudoinverse of Z. The calculation of R is computationally costly, however R only depends on B, which is defined by the parameters of the analysis process, therefore the only operation that is performed at each frame is $\hat{G}_i = RX_i$.

We must note that in contrast to NMF, our gains $\hat{G}_i$ can take negative values. In order to have a proper likelihood we we define the pitch likelihood as:

$$P_i = [\hat{G}_i]_+ / sum([\hat{G}_i]_+) \tag{5}$$

where $[Z]_+$ denotes the operation of setting to 0 all the negative values of a given vector Z.

3.2 Timbre Classification

Estimating the pitch track of the target instrument requires determining when the instrument is not active or not producing a harmonic signal (e.g. in fricative phonemes).

We select a limited number of pitch candidates n_d by finding the largest local maxima of the pitch likelihood function P_i 5. For each candidate a feature vector c is calculated from its harmonic spectral envelope $e_h(f)$ and a classification algorithm predicts the probability of it being a *voiced* envelope of the target instrument. The feature vector c of each of the candidates is classified using Support Vector Machines (SVM). The envelope computation $e_h(f)$ results from the Akima interpolation [1] between the magnitude at harmonic frequencies bins. The timbre features c are a variant of the Mel-Frequency Cepstrum Coefficients (MFCC), where the input spectrum is replaced by an interpolated harmonic spectral envelope $e_h(f)$. This way the spectrum values between the harmonics,

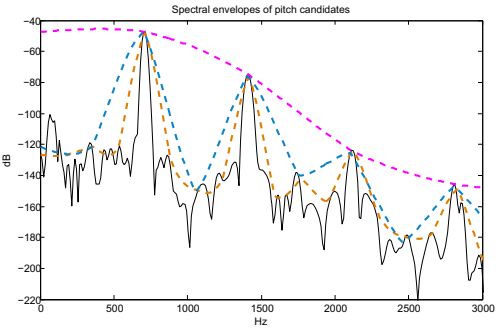

Fig. 1. Spectrum magnitude (solid black line) and the harmonic spectral envelopes (colored dashed lines) of three pitch candidates

where the target instrument is often not predominant, have no influence on the classification task. Figure 1 shows an example of a spectrum $X_i[f]$ (in black) of a singing voice signal, and the interpolated harmonic spectral envelopes $e_{h,1}(f)$, $e_{h,2}(f)$ and $e_{h,3}(f)$ (in magenta, blue and orange respectively), of three different pitch candidates.

The features vector c contains the first 13 coefficients of the Discrete Cosine Transform (DCT), which are computed from the interpolated envelope $e_h(f)$ as:

$$c = DCT\left(10 \cdot \log\left(E[k]\right)\right) \tag{6}$$

where $E[k] = \sum_{f_{k,low}}^{f_{k,high}} e_h(f)^2$, and $f_{k,low}$ and $f_{k,high}$ are the low and high frequencies of the k^{th} band in the Mel scale. We consider 25 Mel bands in a range $[0...5kHz]$. Given an audio signal sampled at $44.1kHz$, we use a window size of 4096 and a hop size of 512 samples. The workflow of our supervised training method is shown in Figure 2. Two classes are defined: *voiced* and *unvoiced* in a frame-based process[1]. *Voiced* frames contain pitched frames from monophonic singing voice recordings (i.e. only a vocal source). Pitched frames have been

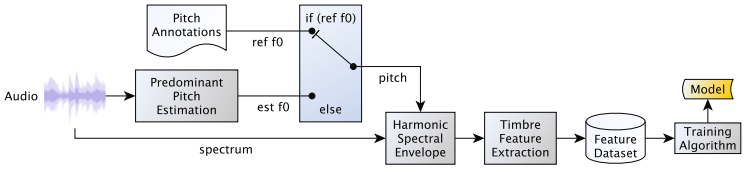

Fig. 2. In the training stage, the $e_h(f)$ is based on the annotated pitch if it exists *if (ref. f0)*, and on the estimated pitch otherwise

[1] The original training and test datasets consist of $384,152$ $(160,779/223,373)$ and $100,047$ $(46,779/53,268)$ instances respectively. Sub-sampled datasets contain $50,000$ and $10,000$ respectively. Values in brackets are given for the voiced and unvoiced instances respectively.

manually annotated. In order to generalize well to real audio mixtures, we also include audio examples composed of an annotated vocal track mixed artificially with background music. *Unvoiced* frames come from three different sources: *a)* non-pitched frames from monophonic singing voice recordings (e.g. fricatives, plosive, aspirations, silences, etc.); *b)* other monophonic instrument recordings (sax, violin, bass, drums); and *c)* polyphonic instrumental recordings not containing vocals. We employ a radial basis function (RBF) kernel for the SVM algorithm [3]. As a pre-process step, we apply standardization to the dataset by subtracting the mean and dividing by the standard deviation. We also perform a random subsampling to reduce model complexity. We obtain an accuracy of 83.54%, when evaluating the model against the test dataset.

3.3 Instrument Pitch Tracking

The instrument pitch tracking step is a dynamic programming algorithm divided into two processes. First a Viterbi is used to find the optimal pitch track in the past C frames, using pitch likelihood P_i for the state probability. Then a second Viterbi allows us to determine the optimal sequence of *voiced* and *unvoiced* frames using the probability found on the timbre classification step for the state. In both cases frequency differences larger than 0.5 semitones between consecutive frames are used to compute transition probabilities. Our implementation works on an online manner with a latency of $C = 20$ frames (232 ms). Due to lack of space the details of the implementation are not presented here.

4 Evaluation

The material used in the evaluation of the source separation method consists of 15 multitrack recordings of song excerpts with vocals, compiled from publicly available resources (MASS[2], SiSEC[3], BSS Oracle[4])

Using the well known BSSEval toolkit [12], we compare the Signal to Distortion Ratio (SDR) error (difference from the ideal binary mask SDR) of several versions of our algorithm and the IMM approach [4]. The evaluation is performed on the "all-minus-vocals" mix versions of the excerpts. Table 1 presents the SDR results averaged over 15 audio files in the dataset. We also plot the results of individual audio examples and the average in Figure 4. *Pan-freq mask* method results in applying the m^{pf} mask from Equation (1). The quality of our low-latency approach to source separation is not as high as for off-line methods such as IMM, which shows an SDR almost 3 dBs higher. However, our LLIS-SVM method shows an increase of 2.2 dBs in the SDR compared to the LLIS-noSVM method. Moreover, adding azimuth information to the multiplicative mask (method *LLIS-SVM-pan*) increases the SDR by 0.7 dBs.

[2] `http://www.mtg.upf.edu/static/mass`
[3] `http://sisec.wiki.irisa.fr/`
[4] `http://bass-db.gforge.inria.fr/bss_oracle/`

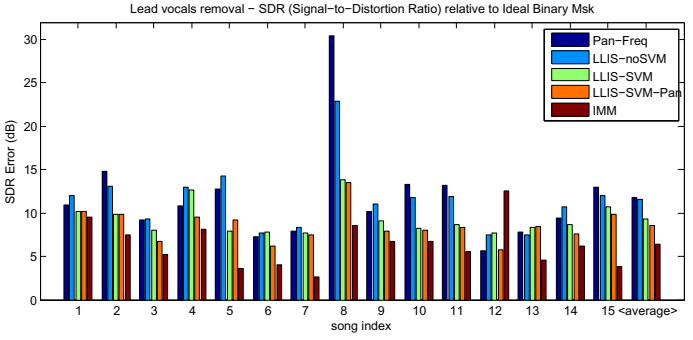

Fig. 3. SDR Error for four methods: pan-frequency mask, LLIS and IMM

Table 1. Signal-To-Distortion Ratio (in dB) for the evaluated methods. The Ideal column shows the results of applying an ideal binary mask with zeros in the bins where the target source is predominant and ones elsewhere.

Method	pan-freq	LLIS-noSVM	LLIS-SVM	LLIS-SVM-pan	IMM	Ideal
SDR-vocals	0.21	0.47	2.70	3.43	6.31	12.00
SDR-accomp	4.79	5.05	7.28	8.01	10.70	16.58

5 Conclusions

We present a source separation approach well suited to low-latency applications. The separation quality of the method is inferior to offline approaches, such as NMF-based algorithms, but it performs significantly better than other existing real-time systems. Maintaining low-latency (232 ms), an implementation of the method runs in real-time on current, consumer-grade computers. The method only targets the harmonic component of a source and therefore does not remove other components such as the unvoiced consonants of the singing voice. Additionally it does not remove the reverberation component of sources. However these are limitations common to other state-of-the-art source separation techniques and are out of the scope of our study.

We propose a method with a simple implementation for low-latency pitch likelihood estimation. It performs joint multipitch estimation, making it well-adapted for polyphonic signals. We also introduce a technique for detecting and tracking a pitched instrument of choice in an online manner by means of a classification algorithm. This study applies the method to the human singing voice, but it is general enough to be extended to other instruments.

Finally, we show how the combination of several sources of information can enhance binary masks in source separation tasks. The results produced by the ideal binary mask show that there are still improvements to be made.

References

1. Akima, H.: A new method of interpolation and smooth curve fitting based on local procedures. JACM 17(4), 589–602 (1970)
2. Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE Transactions on Audio, Speech, and Language Processing 14(1) (2006)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
4. Durrieu, J.L., Richard, G., David, B., Fevotte, C.: Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 564–575 (2010)
5. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Comput. 21, 793–830 (2009)
6. Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.: F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, p. V (May 2006)
7. Goto, M., Hayamizu, S.: A real-time music scene description system: Detecting melody and bass lines in audio signals. Speech Communication (1999)
8. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures. In: Proc (ICASSP) International Conference on Acoustics, Speech, and Signal Processing (2000)
9. Ozerov, A., Vincent, E., Bimbot, F.: A General Modular Framework for Audio Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 33–40. Springer, Heidelberg (2010)
10. Ryynänen, M., Klapuri, A.: Transcription of the singing melody in polyphonic music. In: Proc. 7th International Conference on Music Information Retrieval, Victoria, BC, Canada, pp. 222–227 (October 2006)
11. Sha, F., Saul, L.K.: Real-time pitch determination of one or more voices by nonnegative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1233–1240. MIT Press (2005)
12. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)
13. Vinyes, M., Bonada, J., Loscos, A.: Demixing commercial music productions via human-assisted time-frequency masking. In: Proceedings of Audio Engineering Society 120th Convention (2006)
14. Yeh, C., Roebel, A., Rodet, X.: Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. Trans. Audio, Speech and Lang. Proc. 18, 1116–1126 (2010)
15. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52(7), 1830–1847 (2004)